

Interrogation de sources biomédicales : prise en compte des préférences de l'utilisateur

Sarah Cohen Boulakia*, Christine Froidevaux*, Séverine Lair**

*Laboratoire de Recherche en Informatique, CNRS UMR 8623
Université Paris-Sud, 91405 Orsay Cedex
{cohen, chris}@lri.fr

** Institut Curie, CNRS UMR 144
26 rue d'Ulm, 75248 Paris Cedex 05
severine.lair@curie.fr

Résumé. Nous nous plaçons dans le cadre d'un projet de constitution d'une plate-forme intégrative de données biomédicales pour l'étude génomique des cancers. La plate-forme comporte, entre autres, un certain nombre de scénarios d'analyse qui sont proposés à l'utilisateur. A chaque étape d'un scénario qu'il a choisi de réaliser pour les besoins de son étude, l'utilisateur peut être amené à poser une requête nécessitant d'accéder à différentes sources et il doit alors choisir les sources pertinentes. Nous proposons un guide à l'utilisateur sous forme d'un algorithme de sélection de sources adapté à sa requête et à ses préférences. Pour cela, nous explorons quelques spécificités des banques de données biomédicales et définissons différents critères de préférence utiles pour les biologistes. Nous illustrons notre démarche avec un exemple de requête biomédicale.

1. Introduction

Nous nous plaçons dans le cadre du projet européen HKIS qui vise à la constitution d'une plateforme intégrative de données biologiques et biomédicales pour l'étude génomique des cancers. L'objectif de la plate-forme est de permettre à ses utilisateurs d'analyser leurs résultats d'expérience en les combinant avec d'autres données présentes dans des sources accessibles par le Web. La plateforme offre à chaque utilisateur la possibilité de rapatrier sur son espace de travail un ensemble de banques publiques, le temps d'un traitement. Le besoin de travailler en local répond à une double demande : faire tourner des traitements complexes sur un important volume de données et assurer une confidentialité totale sur les données et les outils manipulés durant les traitements. Cet aspect confidentialité est tout à fait crucial pour les biologistes travaillant sur des données sensibles comme les gènes impliqués dans le cancer du sein par exemple. La plateforme a aussi pour but d'offrir la possibilité d'exporter les résultats obtenus vers les banques privées des utilisateurs. Elle comporte en outre un certain nombre de *scénarios d'analyse*, décrivant différentes méthodologies d'analyse des données, qui sont proposés à l'utilisateur. A chaque étape d'un scénario qu'il a choisi, l'utilisateur peut être amené à poser une requête nécessitant d'accéder à différentes sources. Se pose alors le problème du choix de ces sources.

Dans cet article, nous ne détaillons pas l'architecture de la plate-forme ni ne précisons comment les scénarios d'analyse sont recueillis, puis représentés et implémentés dans celle-ci. Nous nous intéressons ici à la phase d'interrogation et visons à améliorer sa qualité en